未訪問エリアの理解支援のための既訪問スポットに基づく類推情報提示手法

潘　健太† 北山　大輔†

† 工学院大学大学院工学研究科情報学専攻　〒 163–8677 西新宿 1-24-2
E-mail: †em18011@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　近年，観光スポットを決める時に Web 上の観光情報を活用して計画を立てることが多くなっている．しかし，ユーザが多くのエリアから訪問したいエリアを決めた上で，さらに自分のイメージに合う観光スポットを探すのは膨大な時間と労力を必要とする．また，ユーザが未訪問スポットに対して期待と不安を感じる場合がある．本研究では，ユーザの未知なスポットに対する理解を支援するためには，既に訪問したことがある観光スポットの特徴を未訪問スポットにあてはめて理解を支援する類推情報提示を提案する．観光スポット自身の特徴を重視するため，各観光スポットの特徴抽出に，ユーザが入力した観光スポットのすべてのレビュー，対象エリアの観光スポットのすべてのレビューを使用する．また，プロトタイプシステムを構築し，既訪問スポットと未訪問スポットとの類推情報の効果を検証する評価実験を行う．

キーワード　観光スポット，類推，理解支援，レビュー，コサイン類似度，TFIDF，調和平均

1. はじめに

旅行先を決定する時，旅行者は観光スポット検索サイトや観光情報に関連する書籍を見て観光スポットを選び，旅行計画を立てる．しかし，ユーザにとって訪問したいエリアを決定した後，さらにエリア内に数多く存在する観光スポットから，自身のイメージから外れない観光スポットを見つけることは容易ではない．行きたい観光スポットが決まっていない場合ではランキングやおすすめ情報を見て観光スポットを決めることが多くなると考えられる．この時，ユーザが選択した観光スポットに対するイメージが曖昧になるため不安を感じる場合がある．

近年，観光業とソーシャルネットワーキングサービスの発展スピードが加速しており，体験した観光スポットに対するレビューを観光スポット検索サイトに投稿しているユーザが増加している．さまざまな観光スポットを効果的に理解するためには，既存の情報をもとにして，未知な情報と既知な情報との対応関係を考えることが不可欠となる．この考え方は，以前に経験した事柄 (ベースと呼ぶ) を，現在直面している事柄あるいは問題 (ターゲットと呼ぶ) にあてはめる類推に相当する．たとえば，金沢の「にし茶屋街」という未知なスポットに対して既訪問の京都の「花見小路」と似ていると説明するとイメージの理解がしやすくすることがある．

本研究では，ユーザの未知なスポットに対する理解を支援するため，既に訪問したことがある観光スポットの特徴を未訪問スポットにあてはめて理解を支援する類推情報提示を提案する．具体的には，ユーザが入力した既訪問スポットと未訪問エリアから，レビュを用いて既訪問スポット内の各スポットの独特な特徴と未訪問エリア内の各スポットの独特な特徴を抽出し，比較を行って類推情報を提示する．このプロトタイプシステムにより，ユーザが未訪問エリアに対する理解の支援を目指す．図

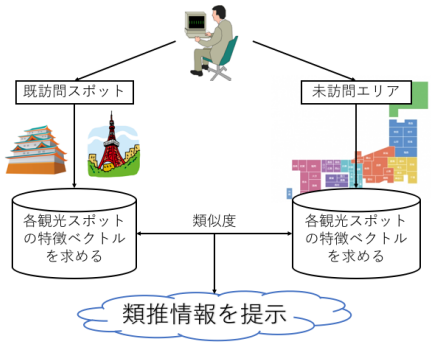

図 1　未訪問エリアの理解支援のための既訪問スポットに基づく類推情報提示手法

1 は提案手法の概念図である．

本論文の構成は下記のとおりである．2 節では関連研究について述べる．3 節では提案手法の概要について述べる．4 節では構築したプロトタイプシステムの効果を検証する評価実験と考察について述べる．最後に 5 節ではまとめと今後の課題について述べる．

2. 関連研究

クチコミを使用した地理情報の検索および推薦に関する研究は数多く行われている．廣嶋ら [1] は，地理情報検索の際のクエリ入力支援として，提示する特徴語の抽出手法について研究を行った．この手法では，各ブログ記事から特徴語候補の抽出および地点の特定を行った．具体的には，特徴語の候補を Wikipedia の見出し語に限定し，ポアソン確率を用いて特徴語抽出を行った．

観光地検索するとき，松本ら [2] はクチコミから特徴語を抽

出して利用する研究を行なった．抽出対象を任意の名詞として，4 種類の手法，TFIDF，ATF(Average Term Frequency)，ポアソン確率，エントロピーのうちどの手法が特徴語抽出に適しているのか検討を行なった．また，抽出した特徴語を利用した検索支援システムを試作し，実験を通して特徴語提示の効果を検証した．

上原ら [3] は Web から観光情報を抽出し，複数の特徴ベクトルから観光地間の類似性を評価することで，観光地を推薦するシステムを提案した．観光地の特徴ベクトルは，知恵袋・ブログ上での共起キーワードと時系列分布，知恵袋上でのカテゴリ構造，観光地周辺施設，地図画像から生成した．これらの特徴ベクトルから観光地間の類似度の測定を行い，類似度の高い観光地を推薦した．

野守ら [4] は日本全国の観光地のクチコミデータを用いて，観光客が話題にする観光テーマを確率的に抽出し，そのテーマを軸として各観光地の特徴を定量的に評価した．また，クチコミのテキストデータにテキストマイニングを実行して表現を抽出し，観光地ごとにその表現の出現頻度を集計したクロス集計表に PLSA を実行することで，観光客のクチコミだけに基づいた観光テーマの抽出と観光地の特徴分析を行なった．

類推に関する研究の数多くは，ベースとなる物語とターゲットとなる問題が与えられ，物語の特徴を問題の特徴にマッピングして問題を解決するものである [5]．石田ら [6] は新知識を理解するための類推能力の育成についての研究を行った．具体名詞を明確な事物，抽象名詞を不明確な事物とし，明確な事物をターゲットとして扱う．類推による新知識の理解の枠組みを実装したシステムを利用することで，類推プロセスの理解と繰り返しによる練習を促し，類推能力を育成できるのかについて検証を行った．砂山ら [7] は自身が知っている知識を，その知識を知らない他人に伝える際に，類推を用いて分かりやすく説明するスキルの獲得を支援するシステムを提案した．評価実験により，システムを用いた被験者が，分かりやすい説明を行える能力を身につけられる可能性を確認した．中村ら [8] は新たな概念を創造しようとするとき，本質的な役割を果たす高次認知機能として類推に着目して研究を行った．構造の類似性には 3 種類あり，特徴の共有数で決まる「対象レベルの類似性」，ベースに存在する関係とターゲットに存在する関係の共有度に基づく「関係レベルの類似性」，および題の解法あるいは目標レベルでの類似性である「プラグマティックな類似性」とがある [9]．

従来のレビューを利用する手法では，クチコミを分析して，対象目標の検索をしやすくための研究が多い．また，類推技術は学習支援でよくて使われている．本研究，提案する未訪問エリアの理解支援のための既訪問スポットに基づく類推情報提示手法は，既訪問スポットと未訪問エリアのレビューを使い，未知ターゲットに対する理解を支援するため，類推の質を明示的扱う．そのため，構造の類似性「関係レベルの類似性」に近いと考えられる．

3. 類推情報提示手法

我々は，未訪問エリアの理解支援のための既訪問スポットに基づく類推情報提示手法を提案する．具体的にはまず，ユーザが既訪問の複数個の観光スポットと訪問したい観光スポットエリア情報を入力する．既訪問スポットレビューベクトルを使って既訪問スポット毎の特徴ベクトルを求める．未訪問スポットも同様にエリア内の各スポットの特徴ベクトルを求める．次に，既訪問スポットレビューベクトルと未訪問スポットレビューベクトルの差分特徴に類似する特徴を持つ未既訪問観光スポット関連付けを行う．最後に，TFIDF を用いて未訪問エリアの理解支援のための類推情報を定義し，ユーザに提示する．

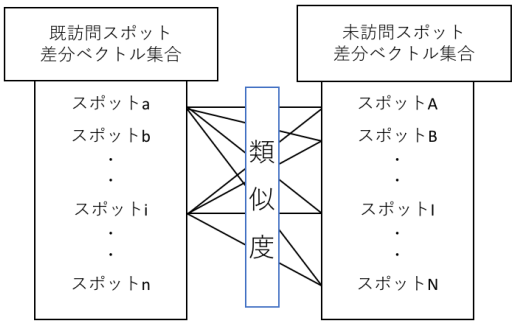

図 2　類似度計算概念図

3.1 観光スポットの相対的な特徴

本研究では，観光スポットの特徴は相対的な特徴を利用する．相対的な特徴とは，特定の観光スポットが，ある観光スポット集合に含まれた他の観光スポットと比較した場合における独特な特徴である．例として，観光スポット集合内に鹿苑寺と清水寺が存在する場合を考える．このとき鹿苑寺の特徴は，金色，金箔，輝き等となり，清水寺の特徴は，舞台，胎内，一望等となる．どちらも京都に存在する寺院であるため，京都や寺院に関連する特徴は独特な特徴として現れることがない．次に，観光スポット集合内に東京都庁舎展望台と鹿苑寺が存在する場合を考える．このとき鹿苑寺の特徴は，金閣寺，お寺，金色，京都等となり，東京都庁舎の特徴は，展望，夜景，新宿等となる．観光スポットのカテゴリーが大きく異なる場合であれば，カテゴリーとしての特徴が現れる．また，スポット自身の特徴を表すことができる．本研究では，あるスポットが集合内の他のスポットと比較するとき，より各スポットの特徴を明らかにできる相対的な特徴に着目して研究を行う．

3.2 コサイン尺度による類似度計算

既訪問スポットや未訪問スポットのレビューベクトルは，形態素解析器「mecab-ipadic-NEologd」(注1)で分かち書き (原型) したレビューを利用して作成する．その後，Doc2Vec(注2)の Distributed Bag-of-Words を利用して，各スポットの全レビューを使って 300 次元で作成したベクトルを使う．本稿に置いて，レビューデータは 2016 年 09 月末までじゃらん (注3)から取得したものを用いる．

(注1)：https://github.com/neologd/mecab-ipadic-neologd/
(注2)：https://radimrehurek.com/gensim/models/doc2vec.html
(注3)：https://www.jalan.net/kankou/

表 1　形態素解析の例

レビュー文書	園内も広く，気分転換に散歩したりするのにちょうどよい．きれいに清掃などもされていて，気分がよいです．
形態素解析	園内 広い 気分転換 散歩 ちょうど よい きれい 清掃 さ れ い 気分 よい

スポット差分ベクトルは式 1 として定義される．スポット差分ベクトルを求めるスポットを除いたスポット集合の各スポットのスポットベクトルの平均値を引いた値となる．$spot_{set} = \{s_1, s_2, \ldots, s_i, \ldots, s_n\}$ は既訪問スポット集合や未訪問スポット集合となっている．また，s_i は集合内のある観光スポットを示している．

$$v_i = s_i - average(spot_{set} - s_i) \tag{1}$$

既訪問スポットの各特徴差分ベクトル v_i と未訪問スポットの各特徴差分ベクトル v_j から，既訪問スポットと未訪問スポット間の相対的な特徴の類似度（図 2）を求める．類似度計算には，コサイン尺度 (式 2) を用いる．

$$cos(v_i, v_j) = \frac{v_{i1}v_{j1} + v_{i2}v_{j2} + \cdots + v_{in}v_{jn}}{\sqrt{v_{i1}^2 + \cdots + v_{in}^2} \times \sqrt{v_{j1}^2 + \cdots + v_{jn}^2}} \tag{2}$$

既訪問スポットの各特徴ベクトルと未訪問エリア内の各特徴ベクトルの類似度が 0.125 以上かつ，類似度が最も高い既訪問スポットと未訪問スポットの関連付けを行う．

3.3　TFIDF による特徴ベクトル生成

観光スポットのレビューはすべて形態素解析器「mecab-ipadic-NEologd」を使用することで，単語抽出処理を行う．しかし，これらを用いて得られた単語は，日本語として成立しない語が含まれており，これらノイズの削除が必要となる．具体的には，助詞，助動詞，連体詞，記号を削除する (表 1)．

節 3.2 で関連付けした既訪問スポットと未訪問スポットの類推情報は単語形式でユーザに提示するため，ある観光スポットのレビュー集合を文書 i とし，i に対する単語 j が出現するスポット集合の出現回数を $TF_{i,j}$，単語 j がスポット集合の文書数を DF_j，スポット集合内の全スポット数を $|D|$ とした時，そのスポットにおける単語の特徴量は，式 3 で定義される．

$$word_{i,j} = TF_{i,j} \times IDF_j \tag{3}$$

$$IDF_j = log(\frac{|D|}{DF_j}) \tag{4}$$

本手法では，既訪問スポットに関して，ユーザが複数個のスポットを入力する．それぞれのスポットの全レビューをまとめて 1 つの文書と見なし，それ以外のスポットの全レビューも文書とみなすことで，式 3，4 によって TFIDF 値を算出し，既訪問スポット毎の特徴ベクトルとする．

未訪問エリアに関して，ユーザがエリアを指定して入力する．エリア内のそれぞれのスポットの全レビューをまとめて 1 つの文書と見なし，それ以外のスポットの全レビューも文書とみなすことで，式 3，4 によって TFIDF 値を算出し，未訪問エリアのスポット毎の特徴ベクトルとする．

表 2　既訪問スポット集合とみ訪問スポット集合

既訪問スポット	未訪問スポット
鹿苑寺（金閣寺）	皇居東御苑
八坂神社	新宿御苑
清水寺	東京都庁舎展望室
龍安寺	浅草寺
伏見稲荷大社	明治神宮

3.4　調和平均による類推情報提示

既訪問スポットから未訪問スポットをイメージするための類推情報は，単語形式でユーザに提示する．節 3.3 で関連付けした既訪問スポットと未訪問スポットの類推情報は，節 3.2 で求めた各スポットの特徴ベクトルによる調和平均を用いて決定する．調和平均とは，逆数の算術平均の逆数である．既訪問スポットのレビュー文書と，未訪問スポットのレビュー文書に，共通して出現する単語を抽出する．抽出した単語のスコアは式 5 によって定義する．$visited_{word}$ と $unvisited_{word}$ は同じ単語がそれぞれ既訪問スポットの TFIDF 値と未訪問スポットの TFIDF 値を示している．単語スコアの値が大きと既訪問スポットと未訪問スポットのそれぞれの TFIDF 値が大きい，つまり単語がそれぞれの文書に置いて重要度が高いことを示している．よって，単語スコアの上位 10 個の単語を類推情報としてユーザに提示する．

$$score = \frac{1}{\frac{1}{2}(\frac{1}{visited_{word}} + \frac{1}{unvisited_{word}})} \tag{5}$$

4.　予 備 実 験

4.1　特徴ベクトルと特徴差分ベクトルの比較実験

節 3.2 では，既訪問スポット集合と未訪問スポット集合それぞれの各スポットベクトルと他のスポットベクトルを使って類似度を計算することで観光スポットの相対的な特徴を求める手法について，妥当性を調査する予備実験を行なった．

4.1.1　実 験 内 容

既訪問スポット集合と未訪問スポット集合それぞれの各スポットベクトルと他のスポットベクトルを使って類似度を計算する場合と，節 3.2 の既訪問スポット集合と未訪問スポット集合それぞれの各スポットベクトルと他のスポットベクトルの差分を使って類似度を計算する場合のスポットの比較を行った．表 2 は，実験で使う既訪問スポット集合と未訪問スポット集合である．既訪問スポット集合は京都の寺院・神社のカテゴリーから 5 つの観光スポットを選んだ．未訪問スポット集合は東京都に存在する観光スポットから複数のカテゴリーに渡って 5 つの観光スポットを選んだ．

4.1.2　実験結果と考察

表 3 は実験結果となっている．特徴ベクトルを利用する場合，既訪問スポットのカテゴリーは寺院・神社であるため，類似する未訪問スポットもカテゴリーによって影響される．未訪問スポットの 5 つのスポットからカテゴリーが一致するスポットが類似スポットになる結果になている．

一方，節 3.2 の提案手法を利用する場合，既訪問スポット集合内の鹿苑寺や龍安寺と他のスポットと比較するとき庭園とい

表 3　特徴ベクトルや特徴差分ベクトルを利用する時の実験結果

特徴ベクトルを利用		特徴差分ベクトルを利用	
既訪問	類似未訪問スポット	既訪問	類似未訪問スポット
鹿苑寺	明治神宮	鹿苑寺	新宿御苑
八坂神社	明治神宮	八坂神社	明治神宮
清水寺	浅草寺	清水寺	浅草寺
龍安寺	明治神宮	龍安寺	皇居東御苑
伏見稲荷大社	明治神宮	伏見稲荷大社	明治神宮

う相対的な特徴を見つけることができる．同様に，未訪問スポット集合内の新宿御苑や皇居東御苑と他のスポットと比較するとき庭園という相対的な特徴を見つけることができる．結果，鹿苑寺と新宿御苑，龍安寺と皇居東御苑を関連づけることができる．また，鹿苑寺の庭園特徴に関して，同じく庭園特徴の新宿御苑と関連ずける理由として鹿苑寺は庭園中に金閣が建ているに対して，新宿御苑は庭園中に旧御涼亭が建ていることから，庭園以外の要素も考量して類似性を求めていることがわかった．よって，既訪問スポット集合と未訪問スポット集合それぞれの各スポットベクトルと他のスポット差分ベクトルの方はより観光スポットの相対的な特徴を求めることができる．

4.2　平均の重みと調和平均の比較実験

節 3.4 では，調和平均を利用して単語スコアを算出し，上位 10 個の単語を類推情報としてユーザに提示するの提案手法について，妥当性を調査する予備実験を行った．

4.2.1　実 験 内 容

調和平均を利用して類推情報を提示する場合と，各スポットの平均の重みを利用する場合の比較を行った．まず，訪問スポットのレビュー文書と，未訪問スポットのレビュー文書に，共通して出現する単語を抽出する．次は，閾値を決めて各スポットの TFIDF 値の平均以上の単語を抽出する．具体的に，既訪問スポット集合の各スポットの TFIDF 値の平均を計算し，スポット毎の単語の TFIDF 値が各スポットの平均の重み以上であるかどうか判断する．また，未訪問スポットも同様な算出方法を使っての集合の各スポットの TFIDF 値の平均を計算し，スポット毎の単語の TFIDF 値が各スポットの平均の重み以上であるかどうか判断する．既訪問スポット単語スコアと未訪問単語スコアの差の絶対値が 0 に近いと，それぞれのスポットにおいての単語の意味合いが近いと示している．よって，差の絶対値が 0 に近い 10 個の単語を類推情報としてユーザに提示する．

4.2.2　実験結果と考察

5.　評 価 実 験

6.　まとめと今後の課題

謝　　辞

文　　献

[1] 廣嶋 伸章, 安田 宜仁, 藤田 尚樹, 片岡 良治, 地理情報検索におけるクエリ入力支援のための特徴語の提示, 第 26 回人工知能学会全国大会, Vol.26, 1C1-R-5-6, 2012
[2] 松本 敦志, 杉本 徹, クチコミから抽出した特徴語を利用する観光地検索支援, 第 75 回全国大会講演論文集, Vol.2013, No.1, pp.307-308, 2013
[3] 上原 尚, 嶋田 和孝, 遠藤 勉, Web 上に混在する観光情報を活用した観光地推薦システム, 社団法人 電子情報通信学会, 信学技報, Vol.112, No.367, pp.13-18, 2012
[4] 野守 耕爾, 神津 友武, 口コミデータに PLSA を適用した観光客目線による観光地分析, 第 29 回人工知能学会全国大会, Vol.29, 1J2-OS-18a-2, pp.1-4, 2015
[5] Gick, M.L. and Holyoak, K.J.: Analogical Problem Solving, Cognitive Psychology, Vol.12, pp.306–355, 1980
[6] 石田 純太, 砂山 渡, 新知識を理解するための類推能力の育成, 第 30 回人工知能学会全国大会, Vol.30, 3F3-3, 2016
[7] 砂山 渡, 石田 純太, 川本 佳代, 西原 陽子, 類推による説明スキルの獲得支援システム, 情報処理学会論文誌, Vol.59, No.10, 1922–1931, 2018
[8] 中村 潤, 大澤 幸生, 概念創造のための類推思考プロセスにおける迷いの効果, 横幹, Vol.2, No.1, p.40-48, 2008
[9] Gentner, D.: Structure-Mapping: A Theoretical Framework for Analogy, Cognitive Science, Vol.7, pp.155–170, 1983
[10] 杉山 将, 確率分布間の距離推定:機械学習分野における最新動向, 日本応用数理学会論文誌,Vol.23, No.3, pp.439-452, 2013